A Metric Framework for quantifying Data Concentration

Peter Mitic[⋆]

Santander UK, 2 Triton Square, Regents Place, London NW1 3AN
Dept. Computer Science, UCL, Gower Street, London WC1E 6BT
Laboratoire d'Excellence sur la Régulation Financière (LabEx ReFi), Paris
`peter.mitic@santandercib.co.uk`

Abstract. *Poor performance of artificial neural nets when applied to credit-related classification problems is investigated and contrasted with logistic regression classification. We propose that artificial neural nets are less successful because of the inherent structure of credit data rather than any particular aspect of the neural net structure. Three metrics are developed to rationalise the result with such data. The metrics exploit the distributional properties of the data to rationalise neural net results. They are used in conjunction with a variant of an established concentration measure that differentiates between class characteristics. The results are contrasted with those obtained using random data, and are compared with results obtained using logistic regression. We find, in general agreement with previous studies, that logistic regressions out-perform neural nets in the majority of cases. An approximate decision criterion is developed in order to explain adverse results.*

Keywords: Copula; Hypersphere; Cluster; Herfindahl-Hirschman; HHI; Credit; Concentration; Decision criterion; Tensorflow; Neural Net

1 Introduction

Successful applications of artificial neural net (hereinafter ANN) methods, and also of other AI methods, are numerous, and particular successes are often reported in the press. A notable recent success in the field of cancer diagnosis is [1]. AI methods have been less successful for credit risk: some credit risk datasets are the 'wrong shape' (the term will be formalised in Section 5). This view is prompted by the following observations:

1. Insensitivity to ANN configuration or tuning
2. Low correlations of single explanatory variables with class
3. Insensitivity to data transformations (e.g. reducing to principal components)

[⋆] The opinions, ideas and approaches expressed or presented are those of the author and do not necessarily reflect Santanders position. The values presented are just illustrations and do not represent Santander data.

4. Insensitivity to attempts to redress the imbalance (e.g. SMOTE, gradient boosting, under-sampling or over-sampling)

Our underlying assumption is that distributional properties of the credit data inhibit prediction of a correct classification. The 'wrong shape' phenomenon is illustrated in Figure 1 which shows two contrasting marginal distributions from two of the data sets considered in this study (see Section 4.1). Data set $LCAB$ with class *credit not approved* on the left shows a loose scatter with no discernable trend or 'shape'. Data set AUS with class *credit approved* on the right shows a concentrated scatter with a trend and a triangular 'shape'. The former type is more typical of credit-related data.

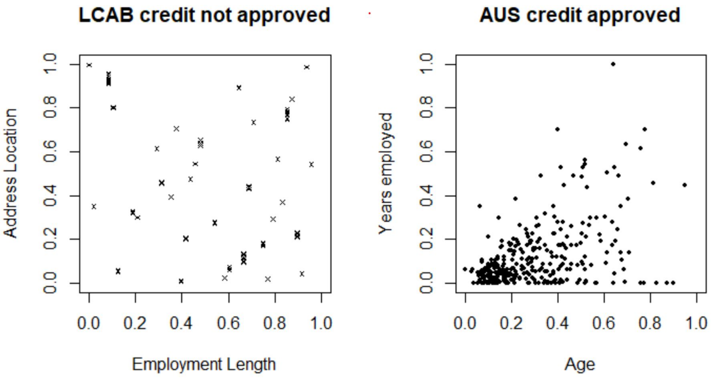

Fig. 1. Marginal Distribution Examples showing contrasting data concentrations

1.1 Economic consequences of credit default

Credit default is very costly for the lender and is a social burden for the borrower and for society. A broad estimate of the amounts involved can be made from UK Regulator figures (https://www.fca.org.uk/data/mortgage-lending-statistics/commentary-june-2019). The outstanding value of all residential mortgage loans at Q1 2019 was £1451bn, of which 0.99% was in arrears. The 2018 capital disclosures from https://www.santander.co.uk/uk/about-santander-uk/investor-relations/santander-uk-group-holdings-plc show that approximately 88% (which is typical) of arrears can be recovered. Therefore the worst case net loss to lenders in the first 3 months of 2019 was $1451 \times 0.99 \times (1 - 0.88) = \pounds 1.724bn$, a very substantial sum!

1.2 Nomenclature and Implementation

In this paper the variable values to be predicted are referred to as *classes*. Typically in the context of credit risk, class determination is a binary decision.

The two classes are usually expressed as categorical variables: 'approved' (alternatively 'pass' or 'good'), and 'not approved' (alternatively 'fail' or 'bad'). Explanatory variables are referred to as *features*. In credit-related data they usually include items such as income, age, address, mean account balance, prior credit history etc. There can be many hundreds of them. The term *tuple* will be used to refer to a single instance of a set of features. Each *tuple* is associated with a single class. The acronyms are are: *LR* for *Logistic Regression* and *AUC* for *Area under Curve*.
The metric calculations were done using the *R* statistical language, and *TensorFlow* was used for neural net calculations. All computations were done using a 16GB RAM i7 Windows processor.

2 Review of Neural Net Applications in Credit Risk

Louzada [2] has an extensive review of the success rate of credit-related applications prior to 2016, using the German and Australian data sets (Section 4.1). The mean success rates of all 30 cases considered were: German: 77.7% and Australian: 88.1%. Those figures are consistently good compared to some we have encountered, but are not comparable to the worst result for the Yala's [1] medical application: 96.2%. More generally, Atiya's pre-2001 review [3] is similar: 81.4% and 85.50% success for two models. Bredart's bankruptcy [4] prediction result is marginally lower: 75.7%.

The results reported by West [5] indicate a general failure of *ANN* methods to improve on results obtained using regressions for the German and Australian data. We used the same data, as well as our own, in Section 4.1 (Section 4.1). We concur with the conclusion that LRs often perform better than AI-based methods: 11.8% greater error rate for *ANN*s. Lessmann [6] gives a lower margin of about 3.2%, using 8 data sets.

There are some better results post-2016. Kvamme et al. [7] reports high accuracy (given as optimal AUC 0.915) using credit data from the Danmarks Nationalbank with a convolutional *ANN*. Addo et al [8], used corporate loan data, and report AUC = 0.975 for their best deep learning model, and 0.841 for their worst. These results are surprisingly good, and we suspect that either the data set used contains some behavioural indicator of default, or that loans in the dataset are only for 'select' customers who have a high probability of non-default. The LC and LCAB data 4.1 have some behavioural indicators (such as amount owing on default, added later), and they are omitted in our analysis. More recently, Munkhdalai et al. citeMunkhdalai2019 reports more relative LR successes: 5.2% better error rate than an *ANN* using a two-stage filter feature selection algorithm, and 7.5% better using a random forest-based feature selection algorithm.

Yampolskiy [10] gives a similar general explanations of AI failure which is particular applicable in the context of credit risk. If a new or unusual situation is encountered in an AI learning process, it will be interpreted, wrongly, as a 'fail' within the context of that process. We suspect that, in the context of assessing credit-worthiness, those new or unusual situations are future events that can only be anticipated with some degree of probability (such as illness, loss of income, mental incapacity).

3 The Concentration Metric Framework

We propose a framwork to measure data concentration, which we think is responsible for the 'wrong shape' phenomenon for credit data. The proposed framework comprises three metrics, each used within a concentration component where the values of the metrics for each class are combined. The idea of a 'framework' is one of extensibility: further metrics can be incorporated in a simple way (see the end of Section 3.1).

3.1 Inter-class Concentration measure

The illustrations in Figure 1 show one instance of a high class concentration and another of low concentration. In order to quantify them, we develop inter-class concentration metrics. Data are partitioned by class, and a concentration metric is calculated for each. They are combined using a variant of an established concentration measure, the Herfindahl-Hirschman Index (*HHI* - see for example [11]). The *HHI* is usually used in economic analysis to measure concentration of production in terms of, for example, percentage of market share or of total sales. We define the index in terms of a metric M_i for class i, associated with a weight w_i (the weight was not part of the original *HHI* formulation). Let M be the sum of the M_i for n classes: $M = \sum_{i=1}^{n} M_i$. Then the *HHI* for metric M is given by $\hat{H}_M$ in Equation 1.

$$\hat{H}_M = \sum_{i=1}^{n} w_i \left(\frac{M_i}{M} \right)^2 \tag{1}$$

In the context of *ANN* classification problems, we use three different interpretations of the metric M_i: M_C, the *Copula* metric M_S, the *Hypersphere* metric, and M_N the *k-Neighbours* metric. The first measures data correlation. The second measures data dispersion and the third measures clustering. For all metrics the weights used (Equation 1) are the proportions of the number of tuples in each class in a training set. The metrics are combined to form the geometric mean concentration measure $\hat{H}$ in Equation 2, which is a general expression for for m metrics. The term *framework* in this paper is used to refer to the applicability of the 'concentration measure + metrics' approach to any required value of m. The geometric mean is used because multiplying the metrics exaggerates the differentiation that each introduces.

$$\hat{H} = \left(\prod_{i}^{m} \hat{H}_i \right)^{\frac{1}{m}} \quad \in (0, 1) \tag{2}$$

In the case of three metrics, Equation 2 reduces to $\hat{H} = (\hat{H}_C \hat{H}_S \hat{H}_N)^{\frac{1}{3}}$.

3.2 The Copula Metric, M_C

A copula is a mechanism for modelling the correlation structure of multivariate data, and thereby generating random samples of any desired distribution. An initial fit to some appropriate distribution is required. Of the common *Elliptic* copulas we choose the multivariate t-copula, as it can capture the effect of extreme values better than the multivariate normal equivalent is able to (see [12] and [13]). Extreme values are often observed in financial return data. It is not necessary to use *Archimedean* copulas, *Clayton*, *Gumbel* or *Frank*, that emphasise extremes even more.

The calculation of the *Copula* metric proceeds by first using a *Fit* function to fit, using maximum likelihood, normal distributions to each of n features data $\{x_i\}$, giving a set of normal parameter pairs $\{\mu_i, \sigma_i\}$. Then we define a t-copula $C_t(c, \nu)$, with $\nu = 3$ degrees of freedom using the covariance matrix c of all the data, and generate a random sample of $m \sim 100000$ U[0,1]-distributed random variables U_i from it using the *R copula* package random number generator, denoted here as $r(C_t)$. The inverse normal distribution function F^{-1} is then applied to the parameter pairs and the values derived from the copula, resulting in a matrix of normal distributions $\{N_i\}$. The row sums of that matrix are then summed to derive the required metric, M_C (Equation 3).

$$\begin{aligned}\{\mu_i, \sigma_i\} &= \{Fit(x_i)\} \\ \{U_i \ &= r(C_t(\nu, c), m\} \\ \{N_i &= F^{-1}(U_i, \mu_i, \sigma_i)\} \\ M_C &= \Sigma(N_i(*, n))\end{aligned} \tag{3}$$

3.3 The Hypersphere Metric, M_S

The *Hypersphere* metric measures the deviation of each tuple that lies within a prescribed hypersphere centred on the centroid of all tuples. For a set of n tuples $t_i, i = 1..n$, denote their centroid by $\bar{t}$, and let the covariance matrix of the set of tuples be c. Then the deviation for tuple t_i is calculated from the Mahanalobis distance, D_i of t_i from $\bar{t}$. The hypersphere refers to the subset of D_i that is within 95% of the maximum of the D_i, and is denoted by $D_i^{(95)}$. The required metric is the sum of the elements of $D_i^{(95)}$ (Equation 4).

$$\begin{aligned}\{D_i\} &= \{\sqrt{(t_i - \bar{t})^T c \ (t_i - \bar{t})}\} \\ D_i^{(95)} &= \{D_i : D_i \leq 0.95 \ max(D_i)\} \\ M_S &= \Sigma_{i=1}^{n} D_i^{(95)}\end{aligned} \tag{4}$$

In practice it makes very little difference if the 95% hypersphere is replaced by, for example, a 90% or a 99% hypersphere.

3.4 The k-Neighbours Metric, M_N

The *k-Neighbours* metric uses a core *k-Nearest Neighbours* calculation. Empirically, we have found that maximal differentiation between classes is achieved by considering the more distant neighbours. Therefore we use the farthest 20% neighbours, not the nearest. The calculation proceeds, for each class, by calculating the Euclidean distances D_i of all the tuples $t_i, i = 1..n$ in each class to the centroid, $\bar{t}$, of that class. The set of distances in excess of the 80^{th} quantile, $Q_{80}(D_i)$ is extracted and summed. We have found that with large datasets, calculating the Mahanalobis distance in place of the Euclidean distance is not always possible due to singularity problems with some covariance matrices. The details are in Equation 5

$$\begin{aligned}\{D_i\} &= \{\sqrt{\Sigma(t_i - \bar{t})^2}\} \\ D_{i,80} &= \{D_i : D_i > q_{80}(D_i)\} \\ M_N &= \Sigma(D_{i,80}) \end{aligned} \tag{5}$$

3.5 Theoetical Metric minimum value

The metric formulations in Equations 1 and 2 admit a theoretical minimum result when using random data with a binary decision. The value of each metric with weights w_i should be $w_i(\frac{1}{2})^2 + (1 - w_i)(\frac{1}{2})^2 = \frac{1}{4}$ (from Equation 1 with $H_1 = H_2$) since random data should yield no useful predictive information. Then, for m metrics, Equation 2 gives the theoretical minimum concentration measure $\hat{H}_{min}$, independent of m in Equation 6

$$\hat{H}_{min} = ((\frac{1}{4})^m)^{\frac{1}{m}} = \frac{1}{4} \tag{6}$$

4 Results

The *ANN* configuration used was: 2 hidden layers with sufficient neurons (always ≤ 100) in each to optimise AUC; typically 100 epochs; ReLU activation in the hidden layers, Sigmoid in the input layer, Softmax in the output layer; categorical cross entropy loss, 66.67% of data used for training.

4.1 Data

Details of the data used in this study are in Table 1. L-Club is the Lending Club (https://www.lendingclub.com/info/download-data.action). UCI is the University of California Irvine Machine Learning database [14]. SBA is the U.S. Small Business Administration. [15]. BVD is Bureau Van Dijk, the Belfirst database (https://www.bvdinfo.com). RAN-P is a randomly generated predictive dataset with two classes, and two highly correlated features. It represents a near minimal concentration with a high predictive element. RAN-NP is similar but is designed to have no predictive element. In all cases, all features are normalised to range [0,1], and there are no missing entries. Where relevant, categorical variables have been replaced by numeric.

Table 1. Data sources

Data	*Source*	*Notes*
INT	Internal	Retail short-term loans
LC	L-Club	All credit grades: LoanStats3b
LCAB	L-Club	Best credit grades A and B only: LoanStats3b
GERMAN	UCI	Statlog German Credit Data
CARD	UCI	Default of credit card clients [16]
AUS	UCI	Statlog Australian Credit Approval
JP	UCI	Japanese Credit Screening
IND	UCI	Qualitative Bankruptcy India
POL5	UCI	Polish Companies Bankruptcy (5-year) [17]
POL1	UCI	Polish Companies Bankruptcy (1-year) [17]
SBA	SBA	'SBA Case' dataset
BVD	BVD	filtered on W. Eur. + Manufacturing Financials
RAN-P	Random	Randomly generated predictive
RAN-NP	Random	Randomly generated non-predictive

4.2 Metric and Concentration results

Table 2 shows the values obtained for the three concentration metrics and the concentration measure (Equations 3, 4, 5 and 2 respectively). The error rates (*Err* columns) are given as proportions, rather than as percentages. It is notice-

Table 2. Distributional Indicators: metrics, $\hat{H}$ and ANN results, in $\hat{H}$ order.

Name	$\hat{H}_C$	$\hat{H}_S$	$\hat{H}_N$	$\hat{H}$	ANN *Err*	ANN AUC	*LR Err*	*LR* AUC
RAN-NP	0.289	0.917	0.885	0.617	0.083	0.540	0.083	0.560
POL1	0.250	0.923	0.911	0.595	0.032	0.590	0.219	0.630
POL5	0.251	0.862	0.877	0.575	0.033	0.62	0.122	0.705
LCAB	0.246	0.769	0.765	0.526	0.330	0.618	0.108	0.635
LC	0.244	0.778	0.606	0.486	0.345	0.679	0.160	0.682
SBA	0.256	0.724	0.521	0.459	0.551	0.680	0.370	0.775
CARD	0.241	0.685	0.470	0.427	0.181	0.775	0.728	0.720
BVD	0.250	0.464	0.426	0.367	0.533	0.870	0.063	0.995
IND	0.374	0.343	0.271	0.326	0.428	0.885	0.012	0.985
GER	0.259	0.350	0.373	0.323	0.245	0.770	0.299	0.820
INT	0.243	0.299	0.349	0.294	0.341	0.815	0.280	0.760
JP	0.253	0.309	0.259	0.273	0.140	0.930	0.252	0.945
AUS	0.249	0.304	0.260	0.270	0.168	0.930	0.342	0.930
RAN-P	0.250	0.262	0.252	0.255	0.305	0.928	0.496	0.680
RAN	0.250	0.250	0.250	0.250	0.501	0.507	0.501	0.506

able from the results in Table 2 that a low $\hat{H}$ value is associated with datasets which work well with ANN processing. Conversely, a high $\hat{H}$ value indicates that ANN processing may not be successful in class determination. LC, LCAB, POL1 and POL5 are the worst cases. The $\hat{H}$ values are more aligned with the AUC values. Figure 2 shows the $\hat{H}$-AUC scatter with a linear trend line ($AUC \sim 1.2 - \hat{H}$, $R^2 = 0.88$), and the $\hat{H}$-Error Rate scatter for comparison. We note that error rate variation with $\hat{H}$ is more volatile than the variation with AUC. Ordinates for the randomly-generated datasets RAN-P and RAN-NP are shown separately. RAN-P represents a borderline *wrong/right shape* boundary and RAN-NP represents a 'worst case' with a minimal predictive element. A further result, not in Table 2 is for randomly generated features with randomly allocated classes (50% in each class). Consistent with Equation 6, we obtained $\hat{H}_C = \hat{H}_S = \hat{H}_N = \hat{H} = 0.25$, with AUC and % success values for ANN and LR all marginally greater than 0.5. Therefore, even 'badly-shaped' datasets are not random!

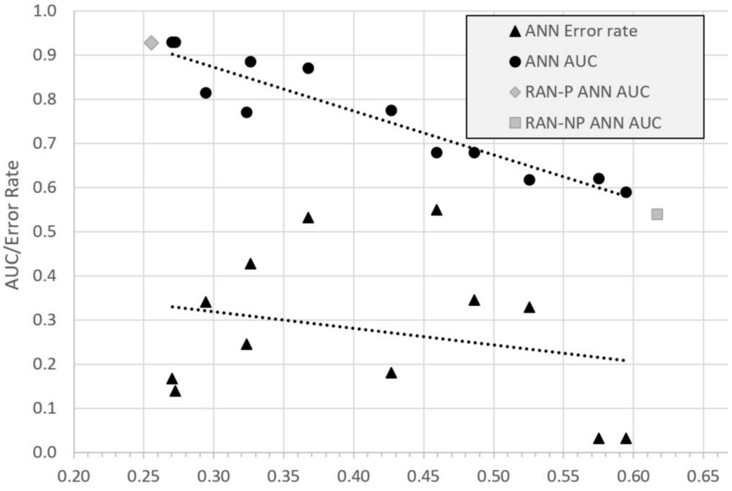

Fig. 2. AUC- and Error rate-Concentration trends.

4.3 Significance Tests

Table 3 shows the results of significance tests for the correlation coefficients for the covariates used to calculate of the two fitted lines in Figure 2 (random data is excluded). The table shows the values of the sum of measured correlation coefficients, r, the calculated t-values and their corresponding p-values. For a theoretical correlation coefficient ρ, with Null hypothesis is $\rho = 0$ and Alternative hypothesis $\rho \neq 0$, the 95% critical t-value is $t_c = 2.228$. The result for the covariate pair $\{ANNAUC/\hat{H}\}$ falls just short of the 95% critical value (at significance level 5.9%).

Table 3. Paired $\hat{H}$ t-test

Covariates	r	t	p
ANN AUC-$\hat{H}$	-0.559	2.13	0.059
ANN Error-$\hat{H}$	-0.347	1.17	0.134

A Sign test on the difference of the *ANN* and LR AUC results (columns *ANN AUC* and *ANN AUC* in Table 2) gives a probability that LR will produce a higher AUC than the *ANN* AUC of 0.0537 (9 cases out of 12): again, just short of a 5% significance level.

5 Discussion

The empirical results in Table 2 give an indication of how the concentration measure $\hat{H}$ can be used to explain any poor results obtained in a *ANN* analysis. Given the result for RAN-P in particular, a decision boundary, $\hat{H}_B$ set at 0.3 is a useful guide. Therefore, a calculated a value of $\hat{H}$, $\hat{H} > \hat{H}_B$ implies that *ANN*-treatment might be unsuccessful or marginally successful (the data are 'wrong'-shaped). Few datasets are successful: {JP and AUS}, and INT is borderline. Dataset RAN-P has been configured specifically to produce a good separation of features so that class can be determined with a high degree of success.

Some characteristics of 'badly-shaped' datasets can be isolated from the metric calculations. A large *Copula* (H_C) metric is often associated with imbalanced data and almost coincident tuples in two or more classes. For example, RAN-NP tuples in class 0 are a random perturbation of its class 1 tuples, corresponding to the {POL1, POL5, LC, LCAB} group. The *Hypersphere* (H_S) metric measures the effect of outliers: either many of them or a smaller number of extremes, or both. Coincident clustering in more than one class is indicated by a high value of the *k-Neighbours* metric M_N.

The value of the concentration metric, $\hat{H}$, should only be seen either as a guide or as an explanatory element of the *ANN* analysis. A high value $\hat{H}$ implies that either the data are too noisy or that they provide insufficient predictive information. When trying to predict credit-worthiness, cases that appear to be high risk sometimes turn out not to be, and vice versa. These cases look like 'noise' in the data, but they are significant because they provide alternative paths to 'success'. It is better to be able to predict a higher proportion of potential credit failures going to deny credit to borrowers who are apparently low risk. Therefore the within-class error rates (i.e. type I and II errors) are also important.

References

1. Yala, A., Lehman, C., Schuster, T., Portnoi, T. and Barzilay, R., A Deep Learning Mammography-based Model for Improved Breast Cancer Risk Prediction, Radiology Online 07/05/2019, https://doi.org/10.1148/radiol.2019182716, 2019

2. Louzada, F., Ara, A. and Fernandes, G.B., Classification methods applied to credit scoring, Surveys in Operations Research and Management Science 21(2):pp117-134, https://doi.org/10.1016/j.sorms.2016.10.001, 2016
3. Atiya, A.F. Bankruptcy Prediction for Credit Risk Using Neural Networks, IEEE Trans. Neural Networks 12(4):pp 929-935, 2001
4. Bredart,X., Bankruptcy Prediction Model Using Neural Networks. Accounting and Finance Research 3(2), 2014
5. West, D., Credit Scoring Models, Computers and Operations Research, 27(11) pp.1131-1152, DOI:10.1016/S0305-0548(99)00149-5, 2000
6. Lessmann,S. , Baesens,B. , Seow,H.. , Thomas,L.C. , Benchmarking state-of-the-art classication algorithms for credit scoring, European Jnl. Operational Research, doi: 10.1016/j.ejor.2015.05.030, 2015
7. Kvamme, H., Sellereite, N., Aas, K. and Sjursen, S., Predicting mortgage default using convolutional works, Expert Systems with Applications 102: pp207-217, https://doi.org/10.1016/j.eswa.2018%.02.029, 2018
8. Addo, P.M., Guegan, D. and Hassani, B. Credit Risk Analysis Using Machine and Deep Learning Models, Risks 6(38), doi:10.3390/risks6020038, 2018
9. Munkhdalai,L., Munkhdalai,T., Namsrai,O., Lee,J.Y. and Ryu,K.H. An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments. Sustainability 11(699), doi:10.3390/su11030699, 2019
10. Yampolskiy, R.V., Predicting future AI failures from historic examples, Foresight, DOI: 10.1108/FS-04-2018-0034, 2018
11. Bikker, J.A. and Haaf, K., Measures of Competition and Concentration in the Banking Industry, Economic and Financial Modelling, 9(2) pp. 53-98, 2002
12. Demarta, S. and McNeil, A.J., The t-Copula and Related Copulas, International Statistical Review 73(1):pp111-129, 2005
13. Rodriguez, C. Measuring financial contagion: A Copula approach. Jnl. Empirical Finance 14(3): pp401-423, https://doi.org/10.1016/j.jempfin.2006.07.002, 2007
14. Dua, D. and Graff, C., UCI Machine Learning Repository Irvine CA, http://archive.ics.uci.edu/ml, 2019
15. Li,M., Mickel,A. and Taylor,S. Should This Loan be Approved or Denied?, Jnl. Statistics Education 26(1):pp 55-66, DOI:10.1080/10691898.2018.1434342, 2018
16. Yeh, I. C. and Lien, C. H., The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. Expert Systems with Applications 36(2): pp2473-2480, 2009
17. Zieba, M., Tomczak, S. K., and Tomczak, J. M. Ensemble Boosted Trees with Synthetic Features Generation in Application to Bankruptcy Prediction. Expert Systems with Applications 58(1); pp93-101, 2016